ELSEVIER

Preventive Veterinary Medicine 45 (2000) 23–41

PREVENTIVE VETERINARY MEDICINE

www.elsevier.nl/locate/prevetmed

Principles and practical application of the receiver-operating characteristic analysis for diagnostic tests

M. Greiner[a,*], D. Pfeiffer[b], R.D. Smith[c]

[a]*Institute for Parasitology and Tropical Veterinary Medicine, Department of Tropical Veterinary Medicine and Epidemiology, Freie Universität Berlin, Königsweg 57, D-14163 Berlin, Germany*
[b]*Department of Farm Animal and Equine Medicine and Surgery, Royal Veterinary College, London, UK*
[c]*Department of Veterinary Pathobiology, College of Veterinary Medicine, University of Illinois, 2001 South Lincoln Avenue, Urbana, IL 61802, USA*

Abstract

We review the principles and practical application of receiver-operating characteristic (ROC) analysis for diagnostic tests. ROC analysis can be used for diagnostic tests with outcomes measured on ordinal, interval or ratio scales. The dependence of the diagnostic sensitivity and specificity on the selected cut-off value must be considered for a full test evaluation and for test comparison. All possible combinations of sensitivity and specificity that can be achieved by changing the test's cut-off value can be summarised using a single parameter; the area under the ROC curve. The ROC technique can also be used to optimise cut-off values with regard to a given prevalence in the target population and cost ratio of false-positive and false-negative results. However, plots of optimisation parameters against the selected cut-off value provide a more-direct method for cut-off selection. Candidates for such optimisation parameters are linear combinations of sensitivity and specificity (with weights selected to reflect the decision-making situation), odds ratio, chance-corrected measures of association (e.g. kappa) and likelihood ratios. We discuss some recent developments in ROC analysis, including meta-analysis of diagnostic tests, correlated ROC curves (paired-sample design) and chance- and prevalence-corrected ROC curves.

Keywords: Receiver-operating characteristic analysis; Diagnostic tests; Likelihood ratios

* Corresponding author. Tel.: +49-30-8108-2317; fax: +49-30-8108-2323.
E-mail address: mgreiner@gmx.net (M. Greiner)

0167-5877/00/$ – see front matter
PII: S0167-5877(00)00115-X

1. Introduction

The crude results of most serodiagnostic tests are measured on ordinal (e.g. grading scheme or sample titration) or continuous (e.g. quantitative readings of single-dilution tests) scales. For all diagnostic tests (except those producing dichotomous outcomes) a value on the original scale is selected as a decision threshold (cut-off value) to define positive and negative test outcomes. Comparison of the dichotomised test results against the true status of individuals (as determined by a reference or ''gold standard'' test) allows estimation of the diagnostic sensitivity (Se, probability of a positive test outcome in a diseased individual) and specificity (Sp, probability of a negative test outcome in a non-diseased individual) (see Greiner and Gardner, 2000). It is well recognised that Se and Sp are inversely related depending on the choice of cut-off value. When increasing values of a measurement are associated with disease, higher (lower) cut-off values are generally associated with lower (higher) Se and a higher (lower) Sp. This relationship has two important implications. First, we would like to select a cut-off value such that the desired operating characteristics (Se, Sp) are achieved. Second, we realise that Se and Sp at a single cut-off value do not describe the test's performance at other potential cut-off values. The latter also implies that the effect of the selected cut-off value should be taken into account when comparing diagnostic tests. These problems are addressed by the receiver-operating characteristic (ROC) analysis and its derivatives.

The ROC methodology was developed in the early 1950s for the analysis of signal detection in technical sciences and was first used in medicine in the late 1960s for the assessment of imaging devices (reviewed by Zweig and Campbell, 1993). ROC analysis has been increasingly used for the evaluation of clinical laboratory tests (Metz, 1978; Henderson, 1993; Schulzer, 1994; Smith, 1995). However, Henderson and Bhayana (1995) reported a lack of consistency with respect to the presentation of ROC analyses. The use of ROC analysis is still limited in the medical and veterinary literature. A systematic review of evaluation (validation) studies of serodiagnostic tests published in 12 biomedical journals in 1995 revealed that ROC analysis has been used in only 3 of 65 medical studies and 1 of 33 veterinary studies (Greiner and Wind, unpublished).

We review practically relevant features of ROC curves and related approaches with emphasis on cut-off selection and test comparison. Data obtained by enzyme-linked immunosorbent assays (ELISAs) for the detection of *Trypanosoma* antibodies will be used as an example. The presentation will refer to continuous ELISA data because this test format is often used for seroepidemiologic applications. The principles, however, apply also to continuous and ordinal diagnostic tests in general. Finally, we describe some extensions of classical ROC-analysis methodology. In the following examples, increasing values of a test result are associated with increasing likelihood of disease.

2. Example data

We use a random subset of data from a validation study of antibody ELISAs for the detection of *Trypanosoma* antibodies in bovine serum. In this study, a negative control group was sampled from non-exposed (Germany) and from exposed (parasitologically

non-infected cattle from a tsetse-infested area in Uganda) cattle populations. The positive control group was sampled from the exposed (parasitologically confirmed) population (Greiner et al., 1997). Test antigen derived from blood-stream form (ELISA A) and in vitro-cultivated procyclic trypanosomes (ELISA B) were evaluated using control sera from 16 exposed non-infected and 4 exposed infected animals. From this experiment we obtained 20 paired optical-density values (OD) (Table 1). In a further experiment, the procyclic ELISA was evaluated using another random selection of 75 non-exposed non-infected and 25 exposed infected animals (ELISA C). In this experiment, the results were expressed as multiples of an internal positive standard (percentage positivity, PP) (Table 2). Preliminary cut-off values had been chosen such that a perfect (100%) Se for ELISA A and B and a perfect Sp for ELISA C were obtained. Cut-off values that fulfil these criteria were 0.86 (Se=1, Sp=0.5), 1.5 (Se=1, Sp=0.5) and 0.7 (Se=0.8, Sp=1) for ELISA A, B and C, respectively. Using ROC analysis and related techniques, we would like (for this example) to rank the three ELISAs according to their diagnostic performance and determine cut-off values that concurrently optimise Se and Sp. Dotplot diagrams for the example data show that there is considerable overlap between negative and positive reference samples (Fig. 1; frequency distribution diagrams are appropriate in case of larger sample sizes). We also note that Se and Sp are a function of the cut-off value. For example, we could achieve a perfect Se (and Sp=0.91) for ELISA C if we were

Table 1
Results of a evaluation study of antibody ELISAs A and B for the detection of *Trypanosoma* antibodies in bovine serum (sub-sample of n=20 from a larger data set described by Greiner et al., 1997)[a]

Animal No.	Reference test[b]	ELISA A	ELISA B
1	0	0.166	0.424
2	0	1.651	2.228
3	0	0.19	0.822
4	0	0.832	1.787
5	0	0.141	0.428
6	0	0.693	1.401
7	1	2.344	2.265
8	0	1.2	0.91
9	0	1.994	2.25
10	0	1.681	2.072
11	0	0.977	1.525
12	0	0.832	1.292
13	0	1.454	1.85
14	0	1.441	1.971
15	1	0.868	1.501
16	0	2.618	1.926
17	0	0.525	0.429
18	0	0.279	0.164
19	1	2.469	2.861
20	1	1.632	2.24

[a] Results expressed as optical-density values (OD). The mean and the standard deviation of 16 reference negative, the mean and the standard deviation of 4 reference test positive animals for ELISA A and B is 1.04, 0.73, 1.83, 0.74 and 1.34, 0.72, 2.22, 0.56, respectively.

[b] Reference test is the parasitological diagnosis of *Trypanosoma* infection (0=negative, 1=positive).

Table 2
Results of a evaluation study of ELISA C for the detection of *Trypanosoma* antibodies in bovine serum of 75 non-infected and 25 infected cattle

ELISA values[a]		
Non-infected (D−)	Non infected (D−)	Infected (D+)
0	0.067	0.254
0	0.075	0.364
0	0.081	0.49
0	0.087	0.509
0	0.095	0.65
0	0.112	0.702
0	0.119	0.716
0.001	0.119	0.743
0.001	0.129	0.752
0.001	0.14	0.879
0.001	0.14	0.899
0.001	0.144	0.927
0.001	0.159	0.937
0.001	0.164	1.057
0.002	0.169	1.064
0.002	0.18	1.081
0.003	0.183	1.116
0.003	0.184	1.263
0.005	0.192	1.346
0.009	0.194	1.402
0.009	0.21	1.665
0.01	0.216	1.698
0.011	0.216	1.799
0.016	0.222	1.801
0.018	0.222	1.934
0.02	0.229	
0.031	0.233	
0.036	0.233	
0.039	0.248	
0.043	0.294	
0.043	0.318	
0.047	0.341	
0.048	0.401	
0.048	0.431	
0.051	0.482	
0.055	0.696	
0.056		
0.058		
0.067		

[a] Results expressed as multiples of the reaction of a positive reference preparation (PP). The mean ($\pm s$) ELISA values of the 75 non-infected and 25 infected animals were 0.11 ($\pm$0.13) and 1.04 ($\pm$0.47), respectively.

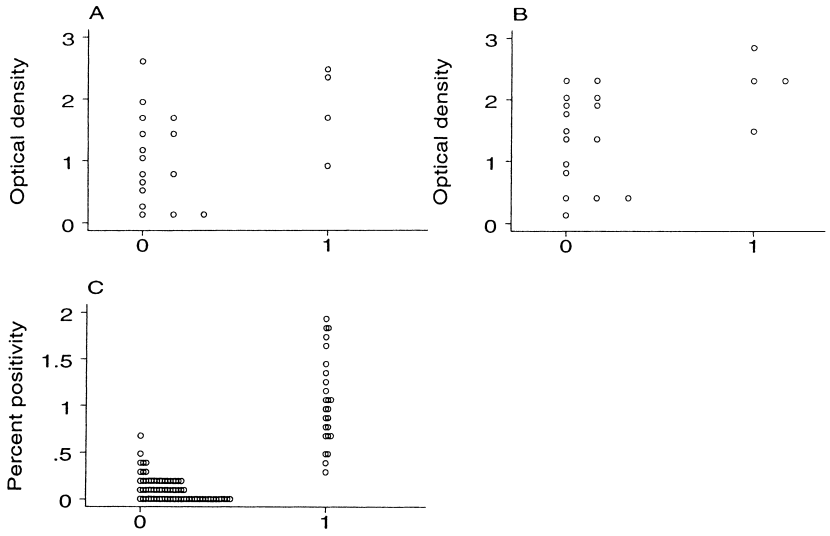

Fig. 1. Dotplots of three *Trypanosoma* ELISAs. ELISA A and B: optical-density values of 16 non-infected (0) and 4 infected (1) animals. ELISA C: PP values of 75 non-infected (0) and 25 infected (1) animals.

to select 0.25 as the cut-off. The example data (pvm_roc.xls) can be downloaded from http://city.vetmed.fu-berlin.de/~mgreiner/pub/data.

3. Basic principles of ROC curves

The underlying assumption of ROC analysis is that a diagnostic variable (e.g. ELISA values) is used to discriminate between two mutually exclusive states of tested animals. During the following discussion, we consider the true disease status (denoted D+ and D− for diseased and non-diseased animals, respectively) but note that various other conditions such as infected/non-infected and protected/non-protected established using an appropriate reference method could also be the aim of diagnostic testing. The statistical distributions of the test values for the D+ and D− subpopulations may be normal (binormal assumption), from different families of distributions (e.g. normal, log-normal) or ordinal.

The diagnostic Se and Sp are a function of the selected cut-off value. ROC analysis assesses the diagnostic performance of the system in terms of Se and (1−Sp) for each possible cut-off value of the test. The terms Se and (1−Sp) in this context are also referred to as ''true-positive fraction'' and ''false-positive fraction'', respectively. Some authors plot the Se against Sp which has no further effect on interpretation because the ROC-curve statistics remain unchanged (e.g. area under the curve) or change in an evident way (e.g. slope).

For tests that yield continuous results (such as ELISA), the cut-off value is shifted systematically over the measurement range and the observed pairs of Se and (1−Sp) are established for each of these k different operating points. The use of a constant, small bin

(interval) width is always valid and recommendable for large data sets. For smaller data sets (or categorical data), it is more economic to use each observed measurement point as a bin limit, which results in variable bin widths for continuous data (also, for categorical data the categories are not necessarily equidistant) but provides an ''optimal resolution''. For sparse data, the shape of the ROC curve cannot be smoothed by increasing the number of bins. For a measurement range of 2, the choice of k=201 would yield a bin width of 0.01. The resulting k pairs {(1−Sp), Se} are plotted in a unit square (i.e. the x and y axes range from 0 to 1). The connection of the points leads to a staircase trace that connects the upper right corner (Se=1, Sp=0 at a cut-off that corresponds to the smallest observed value) to the lower left corner (Se=0, Sp=1 at a cut-off that corresponds to the highest observed value) of the unit square, irrespective of the original unit and range of the diagnostic variable. Note that we assume that the mean value of the negative reference sample is smaller than the mean value of the positive reference sample. Any linear (with negative slope) or inverse transformation of the test data can be used to prepare data for ROC analysis if this assumption does not hold. The empirical trace through the ROC space is referred to as the non-parametric ROC plot. Under the binormal assumption, one can construct a smooth curve as described in the appendix (parametric approach). The latter procedure is termed ''semi-parametric'' if performed on rank-transformed test data (Metz et al., 1998). Detilleux et al. (1999) discuss differences among non-parametric, semi-parametric and parametric methods for ROC analysis using the example of somatic-cell scores for diagnosis of subclinical mastitis. The slope of the smooth ROC curve can be interpreted in terms of the likelihood ratio (LR) of the test (see Appendix B). This relationship will be further discussed below.

ROC curves are invariant with respect to monotone transformations of the original test data such as the linear (with positive slope), logarithmic and square root (Campbell, 1994). In the following discussion, ''*ROC plot*'' denotes a graph of the empirical data whereas ''*ROC curve*'' refers to the smooth ROC function which is an estimate of the true underlying ROC curve using parameters of the empirical sample. ROC plots for the example data are shown in Fig. 2. A number of methods are available for ROC-curve estimation. Tosteson and Begg (1988) used a generalised linear-modelling approach, which allows control of covariates and does not require the assumption of binormality.

Kraemer (1992) and Smith (1995, p. 37) emphasised that Se, Sp and ROC curves are population-specific, despite widespread beliefs advocating the contrary in many text books. True differences in Se and Sp of a test among and within populations and biases in the estimation process of test parameters are described elsewhere (Greiner and Gardner, 2000) and directly apply to ROC analysis (Zweig, 1993).

3.1. Use of ROC analysis to evaluate the discriminatory power of a diagnostic test

The area under the ROC curve (AUC) is a global (i.e. based on all possible cut-off values) summary statistic of diagnostic accuracy. ROC plots for diagnostic tests with perfect discrimination between negative and positive reference samples (no overlap of values of the two groups) pass through the co-ordinates {0;1} which represent 100% Se and Sp. In this case, the AUC would be 1. According to an arbitrary guideline (based on a suggestion by Swets, 1988), one could distinguish between non-informative (AUC=0.5),

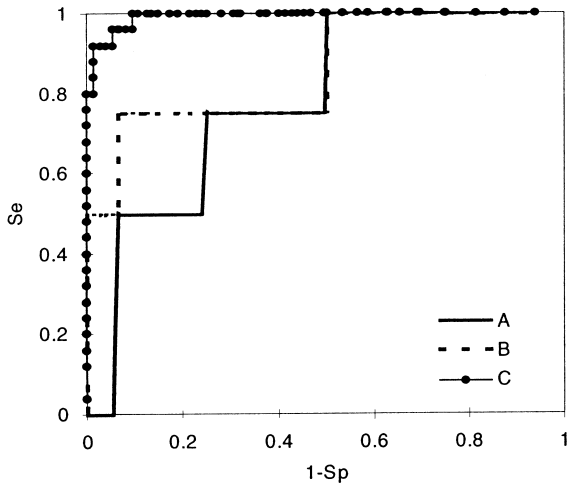

Fig. 2. ROC plots for ELISA A, B and C for the detection of *Trypanosoma* antibodies in cattle.

less accurate ($0.5<AUC\leq0.7$), moderately accurate ($0.7<AUC\leq0.9$), highly accurate ($0.9<AUC<1$) and perfect tests ($AUC=1$). Bamber (1975) and Hanley and McNeil (1982) recognised that the AUC is equivalent to the probability that a randomly drawn individual from the positive reference sample has a greater test value than a randomly drawn individual from the negative reference sample. This interpretation is known as two-alternative forced-choice (2AFC) experiment, where one diseased and one non-diseased individual is presented to the rater (i.e. the diagnostic test) who has to identify the diseased one (see Hanley and McNeil, 1982). Obviously, the probability of a correct answer in a 2AFC experiment (thus the AUC) is not affected by the prevalence in the sample because for each rating the prevalence is fixed at 50% by design.

The AUC summarises the ROC curve as a whole, and therefore attributes the same weighting to both relevant and irrelevant parts of the curve. In practice, one would not select cut-off values from those parts of the ROC curve that have either maximum (lower left part) or minimum slope (upper right part) because other cut-off values exist that lead to better Se without loss of Sp or better Sp without loss of Se, respectively. Furthermore, the diagnostic context might dictate that Se>Sp or Sp>Se or that Se and/or Sp assumes certain minimum values. For example, if the diagnostic context requires that Se is at least 90%, the part of the ROC curve below $Se=0.9$ and its contribution to the AUC statistic is irrelevant for test characterisation and comparison. Detilleux et al. (1999) illustrate the computation of partial areas using the example of somatic-cell score for diagnosis of subclinical mastitis. The AUC statistic gives equal weighting to Se and Sp, which should be considered for interpretation. An appropriate sample estimate of the AUC is the Mann–Whitney *U* statistic in the version of the two-sample rank-sum test. Formulae are shown in Appendix C. This non-parametric interpretation requires no assumptions regarding the distribution of the negative and positive reference samples.

ROC analysis is quite robust to deviations from the binormal assumption (Hanley, 1988). If the test data are approximately normal (or can be transformed into normal) the

distribution of test data can be summarised with the mean values and standard deviations of the two subpopulations. The ROC function can then be parameterised using the standardised mean difference (*A*; which is a measure of the discriminatory power of the test) and the ratio of the two standard deviations (*B*; which is a measure of symmetry). The standard error of the AUC is essential for sample-size calculations (Obuchowski, 1994; Obuchowski and McClish, 1997) and for comparison of ROC curves (see Appendices C–E).

3.2. Use of ROC analysis for the selection of cut-off values

Cut-off values for diagnostic tests can be derived using different methods amongst which the Gaussian (normal) distribution method is most commonly used. Based on this method, a cut-off value is defined as the mean plus two standard deviations (2SD) of the negative reference sample. The rationale of the 2SD procedure is to establish a cut-off value providing an Sp of 97.5% (e.g. Barajas-Rojas et al., 1993). The procedure is clearly not adequate if the test values follow a skewed or multimodal distribution, as is often the case. Moreover, the procedure does not consider the Se; this is the most important disadvantage.

Two parameters (Se and Sp) are necessary to fully describe the probabilities of the four possible test outcomes (TP=true-positive, TN=true-negative, FP=false-positive and FN=false-negative). As described by Schäfer (1989), the cut-off value and the resulting Se (or Sp) can be obtained for a pre-selected Sp (or Se). A plot of Se and Sp as a function of the cut-off value (Fig. 3) provides an useful visualisation and can also be used to derive two cut-off values for the definition of intermediate test results (i.e. test results that are considered non-negative and non-positive) as described elsewhere (Greiner et al., 1995). Optimally, the cut-off selection procedure is an informed decision that takes into account the epidemiologic situation (e.g. prevalence in the target population) and the relative consequences of FN and FP test results (which may differ for every different decision-making situation). As an example, given a disease of low prevalence and high cost of false-positive diagnoses, it may be advisable to choose a cut-off at the lower part of the

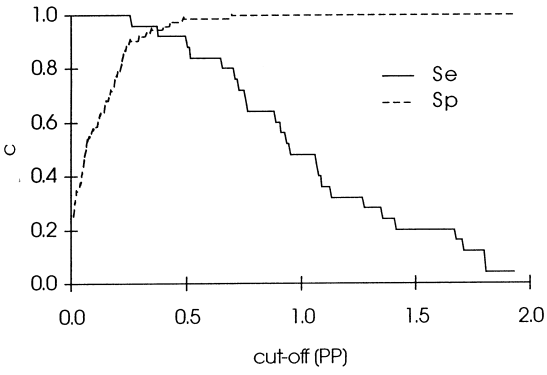

Fig. 3. Plot of the diagnostic sensitivity (Se) and specificity (Sp) of ELISA C as a function of the cut-off value (PP).

curve to maximise Sp. If on the other hand, the disease occurs at high prevalence and missing any diseased animal has serious consequences, a cut-off value towards the upper part of the curve would be selected to maximise Se.

Each application of a diagnostic test is associated with specific consequences of the possible test outcomes. One can attempt to express these consequences (in this context, referred to as ''utilities'') on a common scale (Sondik, 1982; Smith, 1993). Combining ROC analysis with utility-based decision theory can be used to provide an objective, quantitative guide for cut-off selection. This concept is theoretically linked to the ROC curve through the optimality criterion $S=[(1-P)/P][(C_{FP}-C_{TN})/(C_{FN}-C_{TP})]$, where P denotes the prevalence in the target population, C_{FP}, C_{TN}, C_{FN} and C_{TP} represent the utilities associated with the four possible test outcomes, respectively, and S is the slope of the ROC curve at the optimal operating point (Metz, 1978; Smith, 1995, pp. 41f) (Appendix B). The challenge of this approach is that it requires the users to quantify the consequences of each possible test outcome, although outcomes are often thought of only qualitatively. Smith (1995, p. 42) gives an example of Johne's disease for selection of an optimal cut-off value under certain cost and prevalence assumptions.

Generally, without better information, one tends to assume $P=0.5$, $C_{FP}=C_{FN}$, $C_{TN}=C_{TP}$ and a cut-off value would be selected such that $S=1$. The point on the ROC curve closest to the upper left corner of the unit square also optimises prevalence-independent summary measures of Se and Sp such as the Youden index ($J=\text{Se}+\text{Sp}-1$) (Hilden, 1991; measures of diagnostic accuracy are explained in Greiner and Gardner, 2000). Giving equal weights to Se and Sp will very often not only fully exploit the information provided by the diagnostic test in the context of a particular diagnostic objective, but does also facilitate comparison of different diagnostic tests. It implies that the prevalence in the target population is about 50% and that the costs of false-positive and false-negative test results are equivalent. Consequently, this cut-off point might not be optimal for other prevalences and cost ratios.

The slope approach as described above is not a trivial task for empirical (staircase) ROC plots because it requires a smoothed function (e.g. binormal distribution) which introduces additional uncertainties. Therefore, plots of defined optimality criteria as a function of the cut-off value provide a more-practical solution to the problem. Candidates for these criteria are the Se and Sp (Fig. 3), J, efficiency ($\text{Ef}=P\,\text{Se}+(1-P)\,\text{Sp}$) (Fig. 4, top), a misclassification-cost term ($\text{MCT}=(C_{FN}/C_{FP})P(1-\text{Se})+(1-P)(1-\text{Sp})$) (Greiner, 1996) (Fig. 4, middle), odds ratio (OR=antilog [logit (Se)+logit (Sp)]; where the four cells of the 2×2 table are augmented by $\frac{1}{2}$) and the kappa index (Fig. 4, bottom). Since the slope of the ROC curve is equivalent to the LR of the continuous test value, a plot of LR against the cut-off values (Fig. 5, a logarithmic transformation of LR was chosen to improve the readability of the graph) provides an alternative to the plot of MCT. Important aspects for interpretation of these criteria include prevalence-independence (Se, Sp, J, OR, LR), prevalence-dependence (Ef, kappa, MCT, optimised LR for a given prevalence), consideration of misclassification costs (MCT, LR), underlying non-parametric (Se, Sp, J, Ef, OR, MCT, kappa) or logistic regression model (LR), consideration of agreement due to chance (kappa). We suggest that both the non-parametric MCT and logistic regression-based LR are useful for cut-off selection. However, further studies are required to investigate the behaviour of the two criteria for various distributions of test data.

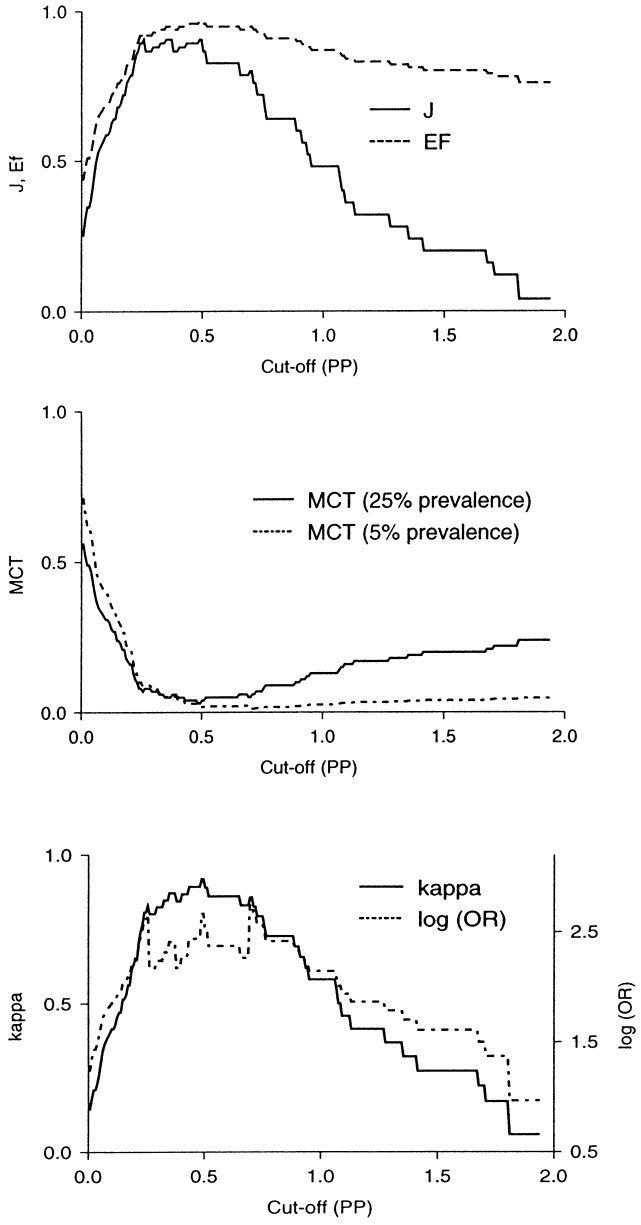

Fig. 4. The efficiency and the Youden index (top), the misclassification cost term for 25 and 5% prevalence (MCT, middle) and the kappa and the log odds ratio (OR, bottom) as a function of the cut-off value for ELISA C. The sample prevalence was 25%.

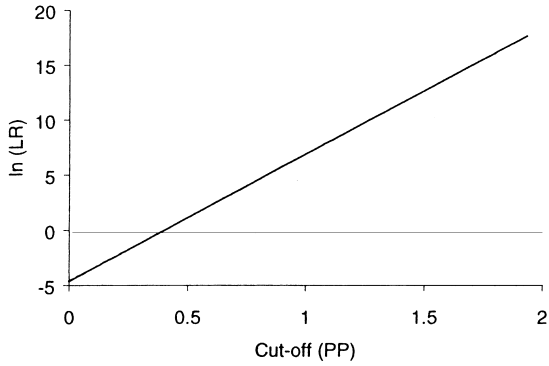

Fig. 5. The natural logarithm of the LR as a function of the cut-off value for ELISA C based on a logistic regression model. The horizontal line marks LR=1 and denotes a cut-off value of 0.4 (PP), which minimised misclassification costs under an assumption of 50% prevalence and equal relative importance of sensitivity and specificity (see Appendix B).

In the ELISA C example, we would select a cut-off value of 0.36 (PP) to achieve maximum accuracy. This result can be read from the plot of Se and Sp as a function of the cut-off value (Fig. 3). Software is available to establish this cut-off value and the corresponding test parameters numerically (Greiner et al., 1995). According to the summary measures Ef and *J*, the MCT, OR and the kappa index, a range of PP between 0.3 and 0.7 would be suitable for the selection of a cut-off (Fig. 4). Algorithms that search for global minima can identify optimal cut-off points (e.g. implemented in CMDT; see Table 3).

Logistic regression analysis is inherently related to the diagnostic-test situation (e.g. Knottnerus, 1992) and can be used for cut-off selection (e.g. Paré et al., 1995). Logistic-regression analysis of the disease state on the continuous test value of ELISA C as described in Appendix B yields the estimated coefficients $a=-5.70$ and $b=11.49$ and under the (arbitrary) optimality criterion $S=\mathrm{LR}=1$, an optimal cut-off value of $X=0.4$ (PP) (see also Fig. 5). An advantage of logistic-regression analysis is that it can be extended to the case of multiple diagnostic tests (Albert, 1982), interaction and dependence (see Hanson et al., 2000). ROC analysis can also be used to evaluate the ''diagnostic discrimination'' of logistic models in general — and, more specifically, in logistic models where the explanatory variables include one or more diagnostic markers and confounders (e.g. age, gender).

3.3. Use of ROC analysis for test comparison

Two or more diagnostic tests may be compared for several reasons. The evaluation of a new test against an established reference test is a special case of test comparison. Often the interest is to compare test parameters (e.g. Se, Sp, LRs, ''overall discriminatory power''; Bennett, 1972; Beam and Wieand, 1991). Such a comparison, however, should account for the effect of the cut-off value as described below. Especially in the context of multiple tests (see Gardner et al., 2000, this issue), one is also interested in the agreement between tests results. For the comparison of the sensitivities (specificities) of two binary

tests, McNemar's test can be used to test the null hypothesis that the sensitivities (specificities) are equal. It is also useful to compare test agreement by calculation of percent agreement and measures of chance-corrected agreement such as kappa for both the diseased and non-diseased populations. The kappa index and chi-square tests can also be used to assess the agreement between two ordinal tests. The kappa index is preferred if an estimation of the degree of agreement rather than a significance test with arbitrary significance level (α) is required. The correlation between paired results of two quantitative tests is sometimes described using the product-moment correlation coefficient (r) based on a linear regression — although measurement errors in both tests invalidate the simple linear model. Graphical (Bland–Altman), rank-based (Passing–Bablok) or modified principal-component (Deming) procedures have been recommended (see Linnet, 1998) to overcome this problem. Plots for visual comparison of tests for a given diagnostic situation (prevalence, misclassification costs) were suggested by Remaley et al. (1999).

As described above, the AUC of a diagnostic test (rather than the Se and Sp at a single cut-off) represents a summary statistic of the overall diagnostic performance of the test. Consequently, AUCs are useful measures for a comparison of the overall diagnostic performance of two tests. However, given the equal weighting attributed to all parts under the curves, it is possible that the comparison of the global AUC will be non-significant for two tests that differ in an area of practical relevance. Comparison of crossing ROC curves may also result in misleading inferences from global AUC estimates. Thompson and Zucchini (1989) describe the use of ANOVA for comparing accuracy indices taking into account these issues.

The non-parametric area under the plots for the example data as established using Eq. (C.1) (and its standard error according to Eq. (C.2)) is 0.781 (0.147), 0.859 (0.125) and 0.993 (0.012) for ELISAs A, B and C, respectively. Approximately (symmetric) 95% confidence intervals (i.e. AUC$\pm$1.96SE (AUC)) overlap for all pairwise comparisons of two tests (data not shown). The pairwise comparison requires an estimate of the correlation (r) between the values of the two tests (see Appendix D). The difference (and 95% confidence interval) according to Eq. (D.1) between the AUCs for the pairwise comparison of A vs B, B vs C and A vs C are 0.078 (-0.029, 0.185; paired sample design, r estimated at 0.85), 0.134 (-0.242, 0.51; two-sample design, r assumed to be 0) and 0.212 (-0.077, 0.5; two-sample design, r assumed to be 0), respectively. The involved sample sizes of non-diseased (n_-) and diseased (n_+) animals for the paired comparison (A:B) was $n_-=16$ and $n_+=4$ and for the unpaired comparisons (A:C, B:C) was $n_-=16$, $n_+=4$ and $n_-=75$, $n_+=25$ for ELISAs A, B and C, respectively. The confidence interval for the differences between the tests includes zero and it can be concluded that despite differences in the AUC estimates, there is no statistically significant difference in performance of the three ELISAs. The question arises about how large a sample size would be required to detect an observed difference of 0.212 (ELISA A vs C) at $p=0.05$. Using A5, we estimate that at least 33 diseased and 33 non-diseased animals should be tested with both ELISAs to confirm that the observed difference of the AUCs is statistically significant. This result is in close agreement with the result ($n=31$) of the AccuROC software (Table 3) and reflects the limitation due to the small sample size (16 non-diseased and 4 diseased animals) that was actually used.

Table 3
Functionality of selected software tools for ROC analysis

Name or trademark	Scope[a]	Availability	Features[b]											Reference (URL location)
AccuROC	ROC	Commercial	1	2	3		5	6	7				11	www.accumetric.com
Analyse-it	Package	Commercial	1	2	3		5		7					www.analyse-it.com
CMDT	Diagn. test	Shareware	1	2	3				7	8	9			city.vetmed.fu-berlin.de/~mgreiner
Episcope	Package	Educational	1	2	3									www.zod.wau.nl/genr/epi.html
GraphROC	ROC	Shareware	1	2	3			6	7		9	10		www.netti.fi/~maxiw/index.html
MedCalc	Package	Commercial	1	2	3		5							www.medcalc.be
NCSS	Package	Commercial	1	2	3									www.ncss.com
PEPI	Package	Shareware	1	2	3									www.usd-inc.com/pepi.html
ROCKIT	ROC	Freeware		2	3	4	5	6	7					www-radiology.uchicago.edu/krl/toppage11.htm
SIMSTAT	Package	Shareware	1											www.simstat.com
Stata	Package[c]	Commercial	1	2	3			6						www.stata.com

[a] Package: statistical software package or add-in, ROC: special ROC analysis software, diagn. test: software for diagnostic test evaluation.

[b] 1: ROC curve on screen, 2: export of ROC curve or scatter plot data to file, 3: area under the curve with standard error or confidence interval, 4: parametric (smoothed) ROC curve, 5: difference between two ROC curves with confidence intervals or standard error, 6: statistical test(s) for ROC curve comparison, 7: paired-sample analysis (correlated ROC curves), 8: resampling techniques, 9: other cut-off functions, 10: partial areas under curves, 11: sample size for ROC curve comparison.

[c] User-defined module by P. Price and F. Wolfe.

4. Recent developments

Confidence bands for ROC curves are needed for inferences from a visual comparison of curves for two or more tests. Methods based on the Greenhouse–Mantel test (Schäfer, 1994), Kolmogorov–Smirnov test and bootstrapping (Campbell, 1994) have been suggested for construction of confidence bands. Confidence intervals for the AUC for diagnostic systems that involve multiple tests were developed by Reiser and Faraggi (1997).

Another topic of current methodological research is the analysis of correlated ROC curves. Correlation between test results must be taken into account if two tests are evaluated using the same set of samples (paired-sample design). Toledano and Gatsonis (1996) suggested an ordinal regression approach and Venkatraman and Begg (1996) used a distribution-free resampling method. Dependence between test results due to repeated measurements of the same animals could be accommodated using generalised estimating equations and the jackknife technique (Beam, 1998) (although generally we would not recommend pooling of repeated measurements).

The ROC approach can also be applied to combine multiple estimates of Se and Sp for one test across several primary evaluation studies. The procedure is known as meta-analysis of diagnostic tests (reviewed by Irwig et al., 1995). One option for a summary measure that takes into account the effect of different cut-off values across the primary studies is the summary ROC function as described by Moses et al. (1983).

Chance-corrected ROC curves have been developed to account for the amount of chance agreement of the conventional parameters Se and Sp (e.g. the QROC concept of Kraemer, 1992; Gefeller and Brenner, 1994) in analogy to the kappa index. In this context, however, there is some controversy about the appropriate derivation of the degree of chance agreement. For example, Holle and Windeler (1997) argued that the suggested chance-correction for Se using (1−Sp) and for Sp using (1−Se) leads to measures that are similar to LRs and that chance-corrected Se, Sp and ROC curves were more difficult to interpret than their conventional counterparts.

5. Software for ROC analysis

Software for ROC analysis is available in various formats including commercial, shareware or stand-alone products, statistical-program packages with built-in or user-defined ROC modules, and spreadsheet calculation macros. Some available programmes are listed in Table 3. However, the list is not comprehensive and we have not compared the relative advantages of the listed programmes. Some features (based on our experience and information provided by the producers) are listed as a guide. A comprehensive evaluation and comparison of the various products would be useful.

6. Conclusions

ROC analysis visualises the cut-off-dependency of ordinal or continuous diagnostic tests and provides an estimate of the accuracy that is independent of specific cut-off

values and prevalence. ROC curves allow a comparison between different diagnostic tests. In addition, the curve provides information which will enable the diagnostician to optimise use of a test through targeted selection of cut-off values for particular diagnostic strategies.

Appendix A. ROC function

The theoretical exponential function that underlies the empirical ROC plot can be estimated under the assumption of normally distributed test values for the non-diseased and diseased individuals (binormal distribution assumption). Let $\bar{x}_0$ and $\bar{x}_1$ (where $\bar{x}_0 < \bar{x}_1$) denote the mean values and s_0 and s_1 denote the standard deviations for the non-diseased and diseased group, respectively. The ROC function is then characterised by the parameter A, which is the standardised mean difference of the responses of the two groups $(A = (\bar{x}_1 - \bar{x}_0)/s_1)$ and the parameter B, which is the ratio of the standard deviations ($B{=}s_0/s_1$). A and B are also referred to as the separation and symmetry parameter, respectively (Metz, 1978). The separation parameter (A) may also be established using s_0 or the pooled standard deviation of the non-diseased and diseased group. The theoretical justification for using s_1 is that under the binormal model, the data can be transformed such that the distributions for non-diseased and diseased individuals are $N(0, 1)$ and $N(\mu, \sigma)$, respectively. Using the normal distribution function (Φ) and the normal deviate (z), we can write the parametric ROC function as

$$\mathrm{Se}(\mathrm{Sp}|A,B) = \Phi\left(\frac{2(1-\mathrm{Sp}) + A}{B}\right).$$

Appendix B. Relationship between ROC curves and likelihood ratios

The slope of the smooth ROC curve (i.e. the tangent at a single point of the function graph) takes values from 0 (upper right corner) to $+\infty$ (lower left corner) and is equivalent to the theoretical LR of the continuous (or ordinal) test value (X) at the respective point of the curve. The LR denotes the ratio of the probability (Pr) of observing the test result in diseased (D+) individuals (Pr(X|D+)) and the probability of observing the same result in non-diseased (D−) individuals (Pr(X|D−)). Because the proportions of D+ and D− individuals with value X may be small in practice, LR cannot be computed as ratio of observed proportions. The appropriate statistical model for LRs for continuous test data is the logistic-regression model: $\mathrm{logit}(\Pr(\mathrm{D}+|\mathrm{X})) = a + bX + \varepsilon$, where a and b are estimated coefficients, ε is the error term and Pr (D+|X) denotes the posterior probability of disease given the test result X. Note that the intercept a depends on the sample prevalence (P'). Using the value x', which denotes the continuous test value that does not change the prior probability of disease ($x' = [\mathrm{logit}(P') - a]/b$), we can define the LR for test value X as

$$\mathrm{LR}(X) = \exp[b(X - x')]$$

(Simel et al., 1993). Using the optimality criterion $S=[(1-P)/P]$ CR, where P and $CR=[(C_{FP}-C_{TN})/(C_{FN}-C_{TP})]$ denote the prevalence in the target population and the cost ratio (see Section 3.2), respectively, we solve $[(1-P)/P]CR=\exp[b(X-x')]$ for X and obtain, consistent with Anderson (1982), a cut-off value for which $LR(X)=S$ as $X=[\text{logit}(P')-\text{logit}(P)+\ln(CR)-a]/b$.

Choi (1998) shows that the LR of a positive (LR+=Se/(1−Sp)) and of a negative (LR−=(1−Se)/Sp) test result at a given point on the ROC curve are equivalent to the slope of the straight line between this point and the lower left and the upper right corner of the ROC square, respectively, and that the LR of a range of test values is equivalent to the slope between the two corresponding points on the ROC curve.

Appendix C. Area under the ROC curve

The most-simplistic approach is to connect the points of the ROC curve with straight lines and to sum the resulting rectangular and triangular areas. This technique (''the trapezoidal rule'') systematically underestimates the true AUC compared with estimates based on a smoothed curve (Vida, 1993). Geometrically, one can show that the trapezoidal AUC is equivalent to $\frac{1}{2}$(Se+Sp) if a single cut-off point is used. The AUC can be estimated with and without assumptions about the distribution of test results. The first, non-parametric approach (sometimes referred to as ''Wilcoxon-area estimate'') is based on the fact that AUC is related to the test statistic U of the two-sample Mann–Whitney rank-sum test (Bamber, 1975; Hanley and McNeil, 1982).

$$\text{AUC} = \frac{n_0 n_1 - U}{n_0 n_1}, \tag{C.1}$$

where n_0 and n_1 (with $n=n_0+n_1$) denote the sample sizes of non-diseased and diseased individuals, respectively, $U = R - \frac{1}{2} n_0(n_0 + 1)$, and R is the rank sum of the negative sample. Under the null hypothesis of a non-informative test, the expected value for the rank sum is $E(R) = \frac{1}{2} n_0(n + 1)$ and therefore, $U = \frac{1}{2}(n_0 n_1)$ and AUC=0.5. The null hypothesis can be assessed using the test statistic $z = (R - E(R))/\sqrt{\text{var}(R)}$, which (for large sample sizes) follows a standard normal distribution. The variance of R can be estimated as $\text{var}(R)=(n_0 n_1 s^2)/n$, where s^2 denotes the sample variance of the combined ranks for both groups. The standard error (SE) of AUC can be derived based on a method described by Hanley and McNeil (1982) and Obuchowski (1994).

$$\text{SE}(\text{AUC}) = \sqrt{\frac{\text{AUC}(1 - \text{AUC}) + (n_1 - 1)(Q_1 - \text{AUC}^2) + (n_0 - 1)(Q_2 - \text{AUC}^2)}{n_0 n_1}}, \tag{C.2}$$

where we use the approximations (Hanley and McNeil, 1982)

$$Q_1 = \frac{\text{AUC}}{(2 - \text{AUC})}, \qquad Q_2 = \frac{2\text{AUC}^2}{(1 + \text{AUC})}.$$

Alternatively, the parametric approach considers the parameters A and B as indicated above and the term $\Phi(z)$ which is the cumulative frequency distribution function of the standard normal distribution (Obuchowski, 1994)

$$\mathrm{AUC} = \Phi\left(\frac{A}{\sqrt{1+B^2}}\right). \tag{C.3}$$

Under the null hypothesis of a non-informative test, we expect $\bar{x}_1 = \bar{x}_0$ (which can be assessed using the two-sample t test). In this case, we get A=0 and AUC=0.5 and the corresponding ROC plot is a diagonal line. The binormal assumption may not be justified for a given set of test data and, therefore, (C.1) is the preferred approach. Maximum-likelihood estimates of the ROC function and the AUC have not been described (Dorfman and Alf, 1968).

Appendix D. Comparison of two ROC curves

Let d denote the difference between the areas under two ROC curves, AUC_1 and AUC_2. Values of d close to zero indicate that the two tests have the same diagnostic performance. We can establish the standard error (SE) of d as

$$\mathrm{SE}(d) = \sqrt{\mathrm{var}(\mathrm{AUC}_1) + \mathrm{var}(\mathrm{AUC}_2) - 2r\,\mathrm{SE}(\mathrm{AUC}_1)\mathrm{SE}(\mathrm{AUC}_2)}, \tag{D.1}$$

where $\mathrm{var}(\mathrm{AUC}_i) = Q_{1i} + Q_{2i} - 2\mathrm{AUC}_i^2$ is an estimate of the variance of the AUC for test i (i=1,2), r is the product-moment correlation coefficient and $\mathrm{SE}(\mathrm{AUC}_i) = \sqrt{\mathrm{var}(\mathrm{AUC}_i)}$ (Hanley and McNeil, 1983) and use 1.96 SE(d) to construct 95% confidence limits around d. In a paired-sample design, r is estimated as the average of the correlation coefficients for the non-diseased and diseased subgroups. For titre results, one could use Kendall's tau instead of r. In a two-sample design, where the two tests have been evaluated using different samples of animals, we set r=0. The covariance adjustment for the paired-sample design (as assumed for ELISA A and B in our example) is generally necessary for the comparison of diagnostic parameters based on the same set of samples (Obuchowski, 1997). Other approaches include an adaptation of the Dorfman–Alf maximum likelihood method (implemented in ROCKIT, Table 3) and resampling techniques (e.g. Venkatraman and Begg, 1996, implemented in CMDT, Table 3).

Appendix E. Sample size

For comparison of two tests with an anticipated difference $d = \mathrm{AUC}_1 - \mathrm{AUC}_2$ involving the same number (n) of diseased and non-diseased animals, significance α and power β, the required sample size is

$$n = \frac{\left(z(\alpha)\sqrt{2\,\mathrm{Var}(\mathrm{AUC}_1)} + z(\beta)\sqrt{\mathrm{Var}(\mathrm{AUC}_1) + \mathrm{Var}(\mathrm{AUC}_2)}\right)^2}{d^2} \tag{E.1}$$

for a one-sided test. For 5% significance and 80% power, we insert $z(\alpha)=1.65$ and $z(\beta)=0.84$, respectively. Obuchowski and McClish (1997) give further detail and describe the case of correlated ROC curves. A comprehensive review of sample-size calculations as well as available software is provided in Obuchowski (1998).

References

Albert, A., 1982. On the use of likelihood ratios in clinical chemistry. Clin. Chem. 28, 1113–1119.

Anderson, J.A., 1982. Logistic regression. In: Krishnaiah, P.R., Kanal, L.N. (Eds.), Handbook of Statistics. North-Holland, New York, pp. 169–191.

Bamber, D., 1975. The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. J. Math. Psychol. 12, 387–415.

Barajas-Rojas, J.A., Riemann, H.P., Franti, C.E., 1993. Notes about determining the cut-off value in enzyme-linked immunosorbent assay (ELISA). Prev. Vet. Med. 15, 231–233.

Beam, C.A., 1998. Analysis of clustered data in receiver operating characteristic studies. Stat. Meth. Med. Res. 7, 324–336.

Beam, C.A., Wieand, H.S., 1991. A statistical method for the comparison of a discrete diagnostic test with several continuous diagnostic tests. Biometrics 47, 907–919.

Bennett, B.M., 1972. On comparisons of sensitivity, specificity, and predictive value of a number of diagnostic procedures. Biometrics 28, 793–800.

Campbell, G., 1994. Advances in statistical methodology for the evaluation of diagnostic and laboratory tests. Stat. Med. 13, 499–508.

Choi, B.C.K., 1998. Slopes of a receiver operating characteristic curve and likelihood ratios for a diagnostic test. Am. J. Epidemiol. 148, 1127–1132.

Detilleux, J., Arendt, J., Lomba, F., Leroy, P., 1999. Methods for estimating areas under receiver-operating characteristic curves: illustration with somatic-cell scores in subclinical intramammary infections. Prev. Vet. Med. 41, 75–88.

Dorfman, D.D., Alf, E., 1968. Maximum likelihood estimation of parameters of signal detection theory — a direct solution. Psychometrika 33, 117–124.

Gardner, I.A., Stryhn, H., Lind, P., Collins, M.T., 2000. Conditional dependence between tests affects the diagnosis and surveillance of animal diseases. Prev. Vet. Med. 45, 107–122.

Gefeller, O., Brenner, H., 1994. How to correct for chance agreement in the estimation of sensitivity and specificity of diagnostic tests. Methods Inf. Med. 33, 180–186.

Greiner, M., 1996. Two-graph receiver operating characteristic (TG-ROC): update version supports optimisation of cut-off values that minimise overall misclassification costs. J. Immunol. Methods 191, 93–94.

Greiner, M., Gardner, I.A., 2000. Epidemiologic issues in the validation of veterinary diagnostic tests. Prev. Vet. Med. 45, 3–22.

Greiner, M., Sohr, D., Gobel, P., 1995. A modified ROC analysis for the selection of cut-off values and the definition of intermediate results of serodiagnostic tests. J. Immunol. Methods 185, 123–132.

Greiner, M., Kumar, S., Kyeswa, C., 1997. Evaluation and comparison of antibody ELISAs for serodiagnosis of bovine trypanosomosis. Vet. Parasitol. 73, 197–205.

Hanley, J.A., 1988. The robustness of the "binormal" assumptions used in fitting ROC curves. Med. Decis. Mak. 8, 197–203.

Hanley, J.A., McNeil, B.J., 1982. The meaning and use of the area under a receiver operating characteristic curve. Radiology 143, 29–36.

Hanley, J.A., McNeil, B.J., 1983. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. Radiology 148, 839–843.

Hanson, T.E., Johnson, W.O., Gardner, I.A., 2000. Log-linear and logistic modeling of dependence among diagnostic tests. Prev. Vet. Med. 45, 123–137.

Henderson, A.R., 1993. Assessing test accuracy and its clinical consequences — a primer for receiver operating characteristic curve analysis. Ann. Clin. Biochem. 30, 521–539.

Henderson, A.R., Bhayana, V., 1995. A modest proposal for the consistent presentation of ROC plots in clinical chemistry. Clin. Chem. 41, 1205–1206.

Hilden, J., 1991. The area under the ROC curve and its competitors. Med. Decis. Mak. 11, 95–101.

Holle, R., Windeler, J., 1997. Is there a gain from chance-corrected measures of diagnostic validity? J. Clin. Epidemiol. 50, 117–120.

Irwig, L., Macaskill, P., Glasziou, P., Fahey, M., 1995. Meta-analytic methods for diagnostic test accuracy. J. Clin. Epidemiol. 48, 119–130.

Knottnerus, J.A., 1992. Application of logistic regression to the analysis of diagnostic data: exact modeling of a probability tree of multiple binary variables. Med. Decis. Mak. 12, 93–108.

Kraemer, H.C., 1992. Evaluating Medical Tests — Objective and Quantitative Guidelines. Sage, Newbury Park.

Linnet, K., 1998. Evaluation of regression procedures for method comparison studies. Clin. Chem. 39, 424–432.

Metz, C.E., 1978. Basic principles of ROC analysis. Semin. Nucl. Med. 8, 283–298.

Metz, C.E., Hermann, B.A., Shen, J.-H., 1998. Maximum likelihood estimation of receiver operating characteristic (ROC) curves from continuously-distributed data. Med. Decis. Mak. 17, 1033–1053.

Moses, L.E., Shapiro, D., Littenberg, B., 1983. Combining independent studies of a diagnostic test into a summary ROC curve — data-analytic approaches and some additional considerations. Stat. Med. 12, 1293–1316.

Obuchowski, N.A., 1994. Computing sample size for receiver operating characteristic studies. Invest. Radiol. 29, 238–243.

Obuchowski, N.A., 1997. Testing for equivalence of diagnostic tests. Am. J. Roentgenol. 168, 13–17.

Obuchowski, N.A., 1998. Sample size calculations in studies of test accuracy. Stat. Meth. Med. Res. 7, 371–392.

Obuchowski, N.A., McClish, D.K., 1997. Sample size determination for diagnostic accuracy studies involving binormal ROC curve indices. Stat. Med. 16, 1529–1542.

Paré, J., Hietala, S.K., Thurmond, M.C., 1995. An enzyme-linked-immunosorbent-assay (ELISA) for serological diagnosis of *Neospora* sp. infection in cattle. J. Vet. Diagn. Invest. 7, 352–359.

Reiser, B., Faraggi, D., 1997. Confidence intervals for the generalised ROC criterion. Biometrics 53, 644–652.

Remaley, A.T., Sampson, M.L., DeLeo, J.M., Remaley, N.A., Farsi, B.D., Zweig, M.H., 1999. Prevalence-value-accuracy plots: a new method for comparing diagnostic tests based on misclassification costs. Clin. Chem. 45, 934–941.

Schäfer, H., 1989. Constructing a cut-off point for a quantitative diagnostic test. Stat. Med. 8, 1381–1391.

Schäfer, H., 1994. Efficient confidence bounds for ROC curves. Stat. Med. 13, 1551–1561.

Schulzer, M., 1994. Diagnostic tests: a statistical review. Muscle Nerve 17, 815–819.

Simel, D.L., Samsa, G.P., Matchar, D.B., 1993. Likelihood ratios for continuous test results — making the clinician's job easier or harder? J. Clin. Epidemiol. 46, 85–93.

Smith, R.D., 1993. Decision-analysis in the evaluation of diagnostic tests. J. Am. Vet. Med. Assoc. 203, 1184–1192.

Smith, R.D., 1995. Evaluation of diagnostic tests. In: Veterinary Clinical Epidemiology. A Problem-Oriented Approach. CRC Press, Boca Raton, FL, pp. 31–43.

Sondik, E.J., 1982. Clinical evaluation of test strategies. A decision analysis of parameter estimation. Clin. Lab. Med. 2, 821–833.

Swets, J.A., 1988. Measuring the accuracy of diagnostic systems. Science 240, 1285–1293.

Thompson, M.L., Zucchini, W., 1989. On the statistical analysis of ROC curves. Stat. Med. 8, 1277–1290.

Toledano, A.Y., Gatsonis, C., 1996. Ordinal regression methodology for ROC curves derived from correlated data. Stat. Med. 15, 1807–1826.

Tosteson, A.N.A., Begg, C.B., 1988. A general regression methodology for ROC curve estimation. Med. Decis. Mak. 8, 204–215.

Venkatraman, E.S., Begg, C.B., 1996. A distribution-free procedure for comparing receiver operating characteristic curves from a paired experiment. Biometrika 83, 835–848.

Vida, S., 1993. A computer program for non-parametric receiver operating characteristic analysis. Comput. Methods Programs Biomed. 40, 95–101.

Zweig, M.H., 1993. ROC plots display test accuracy, but are still limited by the study design. Clin. Chem. 39, 1345–1346.

Zweig, M.H., Campbell, G., 1993. Receiver-operating characteristic (ROC) plots — a fundamental evaluation tool in clinical medicine. Clin. Chem. 39, 561–577.